Comparative physical mapping reveals features of microsynteny between *Glycine max*, *Medicago truncatula*, and *Arabidopsis thaliana*

H.H. Yan, J. Mudge, D.-J. Kim, R.C. Shoemaker, D.R. Cook, and N.D. Young

Abstract: To gain insight into genomic relationships between soybean (*Glycine max*) and *Medicago truncatula*, eight groups of bacterial artificial chromosome (BAC) contigs, together spanning 2.60 million base pairs (Mb) in *G. max* and 1.56 Mb in *M. truncatula*, were compared through high-resolution physical mapping combined with sequence and hybridization analysis of low-copy BAC ends. Cross-hybridization among *G. max* and *M. truncatula* contigs uncovered microsynteny in six of the contig groups and extensive microsynteny in three. Between *G. max* homoeologous (within genome duplicate) contigs, 85% of coding and 75% of noncoding sequences were conserved at the level of cross-hybridization. By contrast, only 29% of sequences were conserved between *G. max* and *M. truncatula*, and some kilobase-scale rearrangements were also observed. Detailed restriction maps were constructed for 11 contigs from the three highly microsyntenic groups, and these maps suggested that sequence order was highly conserved between *G. max* duplicates and generally conserved between *G. max* and *M. truncatula*. One instance of homoeologous BAC contigs in *M. truncatula* was also observed and examined in detail. A sequence similarity search against the *Arabidopsis thaliana* genome sequence identified up to three microsyntenic regions in *A. thaliana* for each of two of the legume BAC contig groups. Together, these results confirm previous predictions of one recent genome-wide duplication in *G. max* and suggest that *M. truncatula* also experienced ancient large-scale genome duplications.

Key words: Glycine max, *Medicago truncatula*, *Arabidopsis thaliana*, conserved microsynteny, genome duplication.

Résumé : Afin de mieux connaître les relations génomiques entre le soja (*Glycine max*) et le *Medicago truncatula*, huit groupes de contigs formés de chromosomes bactériens artificiels (BAC) ont été comparés. Ces groupes couvraient 2,6 millions de paires de bases (Mb) chez le *G. max* et 1,56 Mb chez le *M. truncatula*. Ces comparaisons ont été effectuées suite à une cartographie physique à haute résolution combinée à du séquençage et à des analyses d'hybridation avec des extrémités de BAC correspondant à des séquences à faible nombre de copies. Des hybridations croisées parmi des contigs du *G. max* et du *M. truncatula* ont révélé de la microsynthénie au sein de six des groupes de contigs de même qu'une microsynthénie importante parmi trois groupes de contigs. Au sein des contigs homéologues (régions génomiques dupliquées) chez le *G. max*, 85 % des séquences codantes et 75 % des régions non-codantes étaient suffisamment conservées pour produire une hybridation croisée. Par contre, seuls 29 % des séquences étaient conservées entre le *G. max* et le *M. truncatula* et quelques réarrangements (de quelques kilobases) ont été observés. Des cartes de restriction détaillées ont été produites pour 11 contigs appartenant aux 3 groupes à synthénie élevée et ces cartes suggèrent que l'ordre des séquences est hautement conservé entre régions dupliquées du *G. max* et passablement conservée entre le *G. max* et le *M. truncatula*. Un cas de contigs homéologues chez le *M. truncatula* a également été observé et a été examiné en détail. Une recherche de similarité avec le génome de l'*Arabidopsis thaliana* a permis d'identifier jusqu'à trois régions de microsynthénie chez l'*A. thaliana* pour chacune de deux groupes de contigs des légumineuses. Ensemble ces résultats confirment les prédictions antérieures à l'effet qu'il y aurait eu une duplication génomique récente chez le *G. max* et que le *M. truncatula* aurait aussi connu antérieurement des duplications génomiques à grande échelle.

Mots clés : Glycine max, *Medicago truncatula*, *Arabidopsis thaliana*, microsynthénie conservée, duplication génomique.

[Traduit par la Rédaction]

Received 23 April 2003. Accepted 26 September 2003. Published on the NRC Research Press Web site at http://genome.nrc.ca on 3 February 2004.

Corresponding Editor: F. Belzile.

H.H. Yan and J. Mudge. Department of Plant Pathology, University of Minnesota, St. Paul, MN 55108, U.S.A.
D.-J. Kim and D.R. Cook. Department of Plant Pathology, University of California, Davis, CA 95616, U.S.A.
R.C. Shoemaker. Department of Agronomy and USDA – ARS – Corn Insect and Crop Genetics Research Unit, Iowa State University, Ames, IA 50011, U.S.A.
N.D. Young.[1] Departments of Plant Pathology and Plant Biology, University of Minnesota, St. Paul, MN 55108, U.S.A.

[1]Corresponding author (e-mail: neviny@umn.edu).

 doi: 10.1139/G03-106

Introduction

Comparative genetic, physical, and sequence analysis reveals a striking feature of plant genome evolution: extensive conservation of linkage (synteny) and gene order (collinearity) among related plant species (Gale and Devos 1998; Bennetzen 2000; Keller and Feuillet 2000; O'Neill and Bancroft 2000). Gaut (2002) reassessed some of the published genetic mapping data and estimated that the "synteny probability" (conservation of gene composition without considering relative order) among grass species ranged from 43.4% to 72.6%. Short stretches of microsynteny can even be detected between distantly related species, including *Arabidopsis thaliana* and rice, which have undergone 120–200 million years of divergence (Mayer et al. 2001; Salse et al. 2002).

Comparative sequencing has also revealed that plant genome evolution has involved numerous rearrangements, including translocations and insertions–deletions (indels) (Gaut 2002; Hall et al. 2002). For example, five genes present in a 57-kb region near lateral suppressor in tomato were conserved in *A. thaliana* with only two inversions, and all five of these genes (plus two others missing in tomato) were arranged in the same order and orientation in *A. thaliana* and *Capsella*, with intergenic regions of similar size (Rossberg et al. 2001). Nine of 14 putative genes identified in a 78-kb region near *Adh* of sorghum were conserved in order in a 225-kb region on maize chromosome 1 (Tikhonov et al. 1999), while only four of 33 putative genes predicted in a 340-kb DNA sequence surrounding the *Adh1* and *Adh2* loci on rice chromosome 11 were conserved on maize chromosome 4, excluding *Adh1* (Tarchini et al. 2000).

Segmental and whole-genome duplications are prevalent in the evolution of plant genomes (Shoemaker et al. 1996; Blanc et al. 2000). Analysis of the *A. thaliana* genome sequence indicates that it has undergone three rounds of duplications, with 82% of all genes and 80% of sequences residing in duplicated segments (Simillion et al. 2002). Thus, syntenic relationships are frequently not simply one-to-one but are often one-to-many or many-to-many. In fact, comparisons between the *A. thaliana* genome and the genomes of rice (Mayer et al. 2001), soybean (*Glycine max*) (Grant et al. 2000; Foster-Hartnett et al. 2002), and tomato (Ku et al. 2000) indicate multiple regions forming networks of synteny. For example, a 105-kb region near *ovate* on tomato chromosome 2 contained 17 putative genes, and 12 of these had matches to at least one of four syntenic regions in *A. thaliana* (Ku et al. 2000).

Glycine max is believed to have originated from an ancient polyploid based on several lines of evidence. For example, more than 90% of *G. max* restriction fragment length polymorphism (RFLP) probes detected multiple hybridizing loci, with an average of 2.6 fragments in restriction-digested *G. max* genome DNA (Shoemaker et al. 1996). Marek et al. (2001) found that each *G. max* RFLP probe identifies an average of 2.9 homoeologous regions in *G. max* when probed against two bacterial artificial chromosome (BAC) libraries. However, despite its vast economic importance, it may still be years before *G. max* is fully sequenced due to its intermediate genome size (~1115 million base pairs (Mb)) and relatively complex genome organization. Nonetheless, a deeper understanding of soybean's genome organization and evolution is essential for successful breeding, genetic, and biological research on this globally important crop.

By contrast, *Medicago truncatula* has been proposed as a model to facilitate studies in legume molecular biology and genomics (Cook 1999). *Medicago truncatula* has a relatively compact genome (~470 Mb) with simple diploid genetics, extensive cytogenetic resources (Kulikova et al. 2001), and a funded genome sequencing project that targets the gene-rich portions of the genome (B. Roe, University of Oklahoma, Norman, Okla., personal communication). Examples of substantial microsynteny have been reported between *M. truncatula* and other major legume crops, including *G. max* (Yan et al. 2003), *Pisum sativum* (Gualtieri et al. 2002), and *Medicago sativa* (Endre et al. 2002). For example, the very high level of microsynteny between *M. truncatula* and *M. sativa* has already been exploited to clone a nodulation receptor kinase that is responsible for nod-factor perception in alfalfa (Endre et al. 2002).

Previously, we estimated that 88% of all *G. max* BAC contig groups (defined as a collection of two or more homoeologous BAC contigs identified by the same probe) exhibit some microsynteny, and at least 66% exhibit extensive microsynteny (Yan et al. 2003). Approximately 54% of *G. max* contig groups also show some degree of microsynteny to *M. truncatula* (Yan et al. 2003), although many details remain unclear. Therefore, we have extended our study by comparing eight corresponding genome regions totaling 2.60 Mb in *G. max* and 1.56 Mb in *M. truncatula*, focusing on the features of microsynteny among three groups of *G. max* homoeologous contigs (anchored by *G. max* RFLPs C063, B039, and A685) and their syntenic regions in *M. truncatula*. We compared physical maps between *G. max* homoeologous contigs and between *G. max* and *M. truncatula* and carried out in silico homology searches of sequences sampled from *G. max* – *M. truncatula* contigs against the *A. thaliana* genome sequence. The results suggest high levels of microsynteny (beyond homology) within each of the *G. max* contig groups and also point to a relatively recent duplication event in the *G. max* genome. Moreover, our results indicate one or more ancient duplications in *M. truncatula*. Cases of substantial microsynteny were observed among *G. max*, *M. truncatula*, and *A. thaliana*, but in every case, they were interrupted by both DNA loss and rearrangement.

Materials and methods

Experimental strategy

Mapped *G. max* RFLP clones were hybridized individually to two *G. max* BAC libraries. Positive BAC clones were then assigned to sets of contiguously overlapping clones (contigs) based on fingerprinting results, and many of the BAC ends were sequenced (Marek et al. 2001). Low-copy BAC-end probes developed from one of the contigs (the one with the most sequenced BAC ends) were then hybridized to two *M. truncatula* BAC libraries, and the positive *M. truncatula* BAC clones were fingerprinted, assigned to contigs, and sequenced at both ends. For a given contig, a restriction map was constructed based on a combination of fingerprinting and Southern hybridization results, providing the relative order of *G. max* and *M. truncatula* BAC-end sequences and their intervening physical distances. Finally, all mapped probes

were used in cross-hybridization experiments, allowing us to refine our model for the orientation and extent of conservation among *G. max* and *M. truncatula* contigs.

Isolation of homoeologous BAC contigs in *G. max*

Two *G. max* BAC libraries were screened using mapped *G. max* RFLPs as probes. One library was constructed from *Hin*dIII partially digested genomic DNA of the variety "Williams 82" (with prefix "IS") (Marek and Shoemaker 1997) and the other from *Eco*RI partially digested genomic DNA of the variety 'Faribault' (with prefix "UM") (Danesh et al. 1998). These two libraries together represent 10-fold coverage of the *G. max* genome. DNA of positive BAC clones was prepared using the alkaline-lysis method (Marra et al. 1997), digested with *Eco*RI (IS BACs were also digested with *Hin*dIII), and transferred to nylon membranes. These BAC clones were confirmed by hybridizing filters with the original RFLP probes and then assigned to contigs by fingerprinting (Marra et al. 1997) and using Southern hybridization results. When an RFLP probe detected only one *G. max* contig, BAC-end probes were developed from the contig and used to identify potential homoeologous contig(s). The contig with the most BAC-end sequences was termed "contig 1"; other homoeologous contigs were termed "contig 2", "contig 3", and so on in arbitrary order.

Identification of syntenic BAC contigs in *M. truncatula*

To identify syntenic regions in the *M. truncatula* genome, all low-copy BAC-end probes from each of *G. max* contig 1 (Table 1) plus the original RFLP probe were hybridized to two *M. truncatula* libraries, both constructed from *Hin*dIII partially digested genomic DNA of 'Jemalong' (Nam et al. 1999; D.R. Cook et al., unpublished). In each library screening, DNA of 8–13 low-copy probes from each *G. max* contig were mixed, labeled with ^{32}P, and hybridized to one set of *M. truncatula* library filters containing the entire 30 720 clones of the original *M. truncatula* BAC library (Nam et al. 1999) plus 36 864 (out of 103 683) clones from a second library (D.R. Cook et al., unpublished). Positive *M. truncatula* BAC clones were digested with *Hin*dIII, electrophoresed, and transferred to nylon membranes. The authenticity of these *M. truncatula* BAC clones was confirmed by hybridizing filters containing these BAC clones with a mixture of the *G. max* probes originally used to detect the *M. truncatula* BACs. BAC clones were assigned to tentative contigs, with BACs in each group sharing the same sizes of hybridized fragments. To further clarify which *G. max* probe(s) hybridized to which *M. truncatula* group, polymerase chain reaction (PCR) products corresponding to each of the *G. max* probes were subjected to gel electrophoresis, blotted to nylon membranes, and hybridized with *Hae*III-digested *M. truncatula* BAC clones representing each of the tentative *M. truncatula* contigs. Fingerprinting data were used to build *M. truncatula* contigs as described below. To aid in the identification of *M. truncatula* contigs showing extensive conservation in *G. max*, the above filters containing *Hin*dIII-digested *M. truncatula* BAC clones were also hybridized with two *Hae*III-digested *G. max* BAC clones that cover the entire corresponding *G. max* contig 1.

Sequence analysis

Both ends of *G. max* and *M. truncatula* BAC clones were sequenced by the University of Minnesota Advanced Genetic Analysis Center on ABI PRISM 377 sequencers (Marek et al. 2001). BAC-end sequences (and the corresponding BAC-end probes) were named by a four-digit number followed by "R" indicating right end or "F" indicating left end. These BAC-end sequences, as well as RFLP probe sequences retrieved from GenBank, were first analyzed for similarity using Sequencher™ 3.1.1 (Gene Codes Co., Ann Arbor, Mich.). This analysis identified sequences that were identical (or overlapping) in the same *G. max* or *M. truncatula* contig as well as sequences representing homologs between contigs. BAC-end and RFLP probe sequences were then analyzed against (*i*) the nonredundant GenBank protein database using BLASTx (Altschul et al. 1990) and (*ii*) the *G. max* Gene Index (version 8.0, release date 1 June 2002) and *M. truncatula* Gene Index (version 5.0, release date 1 June 2002) at the Institute of Genome Research (www.tigr.org/tdb/tgi) using BLASTn. In this article, potential coding sequences are defined as those that had significant matches in BLASTx (excluding alignments with plant repetitive DNA) with an expected value of $\leq 1 \times 10^{-8}$ or showed ≥95% identity over at least 100 base pairs (bp) in BLASTn. All remaining sequences were considered to be noncoding.

All *G. max* and *M. truncatula* sequences were further compared with the *A. thaliana* coding sequences (-introns, -UTRs) DNA (http://www.arabidopsis.org/). A cutoff of $\leq 1 \times 10^{-8}$ was considered significant. To rule out complexities caused by members of gene families, sequences that had more than five significant hits were excluded. Cases of microsynteny were inferred when homologs of at least two sequences from a single *G. max* or *M. truncatula* contig and each consisting of distinct predicted open reading frames were less than 200 kb apart in the *A. thaliana* genome.

Isolation of BAC-end probes

Primers were designed for *G. max* and *M. truncatula* BAC-end sequences, excluding those that showed significant homology to plant repetitive DNA sequences. PCR products, 132–558 bp in length, were amplified from DNA of the corresponding BAC clones and purified using either the Qiaquick gel extraction kit or the Qiaquick PCR purification kit (QIAGEN Inc., Valencia, Calif.). Labeled PCR products were hybridized against *G. max* and *M. truncatula* genomic DNA to detect the possible presence of repetitive DNA. Only low-copy probes were used in BAC library screening and cross-hybridization. All BAC-end sequences have been submitted to GenBank, and detailed information about each sequence can be found at http://umn.edu/home/neviny.

Fingerprinting

To build BAC contigs, *G. max* and *M. truncatula* BAC clones identified by the same probes were subjected to agarose gel based fingerprinting (Marra et al. 1997) and were assigned to contigs based on previously described criteria (Yan et al. 2003). Briefly, approximately 50 ng of BAC DNA was digested with *Eco*RI (*G. max* BACs) or *Hin*dIII (*M. truncatula* BACs) at 37 °C for 3–4 h, separated on a 1.0% agarose gel (2× GGB buffer; Wong et al. 1997) at 12 °C for 15 h, and stained for 1 h with Sybr Gold (Molecular Probes, Eugene, Oreg.) diluted to 1 : 10 000 in 2× GGB buffer. Images were acquired by photography with Polaroid

Table 1. Summary of soybean contigs anchored by RFLP probes.

RFLP probe	Gm contig	Linkage group	No. of BACs	Contig size (kb)	Total no. of sequences	Repetitive/ multicopy sequences[a]	Sequence density (kb)[b]
Bng154	Contig 1	H	9	166	14	0	11.9
A598	Contig 1	B1	8	140.5	15	4	12.8
A656	Contig 1	?	7	207.3	14	5	23
B172	Contig 1	A1	8	195.9	15	1	14
A572	Contig 1	?	7	182.3	11	1	18.2
C063	Contig 1	D1a+Q	13	192.2	23	0	8.4
A685	Contig 1	B2	6	181.4	12	0	15.1
B039	Contig 1	I	10	185	17	0	8.4
Total				1450.6	121		

Note: Gm, *Glycine max*; Mt, *Medicago truncatula*; na, not available.
[a]Gm sequences from contig 1 that showed significant similarity to known plant repetitive DNA or producing hybridization signals typical for repetitive or multicopy sequences.
[b]Average physical distance between adjacent low-copy sequences.
[c]No contig with microsynteny was identified in soybean by using low-copy BAC-end and RFLP probes from contig 1 of RFLP probes A572 and B172.
[d]Low-copy BAC-end and RFLP sequences from Gm contig 1.
[e]No contig with microsynteny was identified in Mt by using low-copy BAC-end and RFLP probes from contig 1 of RFLP probes A598 and A656.
[f]No sequence from soybean contig 1 of RFLP B039 was conserved in Mt contig 2 that was microsyntenic to Mt contig 1.
[g]Hybridized by two overlapping soybean BAC clones that spanned the entire Gm contig 1.

instant films and by scanning gels on a FluorImager STORM 840 (Amersham, Amersham, U.K.) in the Imaging Center, University of Minnesota. Fingerprinting gels were then blotted to nylon membranes. Since some of the *G. max* BAC clones were derived from *Hin*dIII partially digested genomic DNA but were digested with *Eco*RI in fingerprinting, these filters were also hybridized with ^{32}P-labeled vector DNA to visualize vector and vector-insert junction fragments. Restriction fragments were manually scored and aligned using the scanned gel images and images captured on Polaroid films along with Southern hybridization results from vector and BAC-end probes (see below).

Cross-hybridization and restriction mapping

To estimate the extent of conservation between different *G. max* and *M. truncatula* contigs in the same contig group, all low-copy BAC-end probes from one contig were hybridized individually to restriction-digested BACs belonging to the other contigs in the same contig group (Table 1). If two or more unique probes showed cross-hybridization between a pair of contigs, an instance of microsynteny was inferred.

For each contig, a restriction map was constructed in two steps. First, each BAC-end probe was hybridized to restriction-digested BAC clones from the same contig to determine the relative order of BAC ends and the fragments containing these BAC ends. Then, fingerprinting and cross-hybridization data that indicated which BACs shared a specific restriction fragment were used to locate each of the remaining restriction fragments to an interval demarcated by adjacent BAC ends. In many cases, the relative order of multiple fragments localized to the same interval could not be determined.

Throughout the experiments, hybridization was conducted at 65 °C overnight. Filters were washed at 65 °C in 2× saline sodium citrate (SSC) – 0.1% sodium dodecyl sulfate (SDS), 1× SSC – 0.1% SDS, and 0.5× SSC – 0.1% SDS for 15 min each. However, library filters were washed only twice in 2× SSC – 0.1% SDS for 10 min each for easy localization of positive BAC clones.

Results

Microsynteny among homoeologous *G. max* contigs

Eight *G. max* RFLP probes were used to screen two *G. max* BAC libraries. Seven probes identified two or three homoeologous contigs each (although the second A572 contig showed no cross-hybridization with the first and was not pursued further), while probe B172 identified just one (Table 1). Restriction maps were constructed for contig 1 for each of the RFLPs, with between five and 13 BAC clones per contig. These contigs ranged in size from 140 to 207 kb, and between 11 and 23 sequence tags (BAC ends and RFLP probes) were physically linked to the constituent restriction fragments. Based on strong similarity to repetitive DNA sequences from other plant species and Southern hybridization patterns on *G. max* genomic DNA, four contigs had one or more BAC ends classified as containing some repetitive and (or) multicopy DNA. In the other four contigs, all BAC ends contained only low-copy sequences (Table 1).

Conservation among contigs identified by the same RFLP probe (a so-called "contig group") was investigated by cross-hybridization using low-copy BAC ends as probes. Microsynteny was detected in six of the seven groups of homoeologous contigs in *G. max*, with between two and 15 BAC-end sequences from contig 1 conserved in one or more homoeologous contig. As both A572, whose contig 1 BAC ends showed no cross-hybridization with a second A572 contig, and B172, which uncovered only a single contig, did not identify microsyntenic *G. max* contigs, all nine low-copy BAC-end probes from A572 contig 1 and 13 probes from B172 contig 1 were used to rescreen the *G. max* BAC libraries. In both cases, additional contigs were identified, but only one sequence from contig 1 was found to be conserved. Thus, we conclude that these two genome regions do not

Microsyntenic contigs in Gm[c]	Size of Gm contigs (kb)	Sequences conserved in Gm[d]	Microsyntenic contigs in Mt[e]	Size of Mt contigs (kb)	Gm sequences conserved in Mt[d,f]	No. of cross-hybridized bands in Mt contig 1[g]
Contigs 2 and 3	106.7, 96.3	2, 4	Contig 1	192.6	2	5
Contig 2	104.6	3	na	na	na	1
Contigs 2 and 3	129, 134.5	4, 5	na	na	na	2
na	na	na	Contig 1	129.7	2	4
na	na	na	Contig 1	196.1	2	2
Contig 2	167.4	15	Contigs 1–3	180.1, 196.7, 130.1	5, 2, 2	9
Contigs 2 and 3	146.3, 119.5	6, 2	Contig 1	159	3	9
Contig 2	146.4 1150.7	11	Contigs 1 and 2	193.7, 183.3 1561.3	6, 0	8

have readily discernable homoeologs elsewhere in the *G. max* genome.

Identification of *M. truncatula* contigs showing microsynteny to *G. max*

To identify syntenic regions in the *M. truncatula* genome, all available low-copy BAC ends plus the original RFLP from contig 1 of the eight *G. max* contig groups were used to screen two *M. truncatula* BAC libraries. For six of the contig groups, between one and three *M. truncatula* contigs with putative microsynteny to *G. max* were identified, each with two to six cross-hybridizing *G. max* probes (Table 1). For each contig group, we also hybridized two *Hae*III-digested *G. max* BAC clones, selected to completely span the underlying contig 1, to filters containing the corresponding *Hin*dIII-digested *M. truncatula* BACs. In three of the contig groups, anchored by RFLP C063, B039, or A685, the two *G. max* BAC clones hybridized to eight or more fragments in a single *M. truncatula* contig. Based on these cross-hybridization results using *G. max* BAC ends and whole BAC clones as probes, we chose to focus on these three contig groups, which apparently represented regions with substantial microsynteny between the *G. max* and *M. truncatula* genomes.

Analysis of BAC-end and RFLP sequences

Eighty *G. max* and 67 *M. truncatula* BAC ends from three contig groups (C063, B039, and A685) were tested by Southern hybridization with *G. max* and *M. truncatula* genomic DNA. Only two *G. max* and two *M. truncatula* BAC ends produced signals suggesting the presence of repetitive DNA. Two other *M. truncatula* BAC ends matched a putative retroelement polyprotein in *A. thaliana*. After excluding these six BAC ends plus four other *M. truncatula* BAC ends that were homologs of *G. max* BAC ends, 137 distinct BAC ends and two RFLP probes were retained for further analysis (RFLP C063 was not included because it is a paralog of BAC-end 7325F). Fifty-nine sequences were classified as coding, showing significant alignments to entries in expressed sequence tag and (or) protein databases.

Restriction maps were constructed for 11 of the contigs in these three contig groups. Maps were constructed for two *G. max* contigs and one *M. truncatula* contig in contig groups A685 and B039 and for two *G. max* contigs and three *M. truncatula* contigs in contig group C063. All BAC ends and RFLP sequences could be physically linked to individual fragments on the restriction maps (Figs. 1–3). Cross-hybridization using BAC ends and RFLP clones as probes revealed that 13 out of 18, 16 out of 22, and 21 out of 23 sequences were conserved between pairs of homoeologous *G. max* contigs anchored by A685, B039, and C063, respectively. The relative order could be determined for a total of 20 pairs of sequences, and every one showed complete conservation (collinearity) in *G. max*. The remaining conserved probes hybridized to regions on syntenic contigs that also suggested conservation of sequence order, although the exact order within individual restriction fragments could not be determined. Between *G. max* and *M. truncatula*, nine out of 32, 17 out of 44, and 15 out of 63 sequences were conserved for the A685, B039, and C063 contig groups, respectively. Sequence order was largely conserved, but a few cases of rearrangements were observed (Figs. 1–3). Details about each of these three contig groups are provided below.

Microsyntenic relationships within the RFLP A685 contig group

Soybean RFLP A685, a putative noncoding sequence without significant BLASTx or BLASTn hits, identified three contigs in *G. max* (Table 1). Restriction maps were constructed for contig 1 (on molecular linkage group B2) and contig 2, each containing six BAC clones. Of 21 sequences isolated from the two contigs, 13 were conserved between the contigs based on cross-hybridization and sequence alignment (Fig. 1*a*). The relative order was conserved for all four pairs of paralogs that could be determined. For the remaining nine conserved sequences whose relative order could not be determined

unambiguously, neighboring probes hybridized to the same restriction fragments in both contigs, reflecting overall conservation of sequence order. Of the eight probes that did not cross-hybridize, three (5474R, 5110R, and 5122R) mapped to the upper end of contig 1 and were presumably located beyond the region of overlap. In contrast with the high conservation observed between contigs 1 and 2, *G. max* contig 3 was far less conserved (not shown). Except for RFLP A685 itself, the only other sequence conserved in contig 3 was BAC-end 5124R from contig 1.

The 12 *G. max* probes from contig 1 together identified 11 *M. truncatula* BACs, nine of which were confirmed in Southern hybridization, fingerprinted, and assigned to a single contig (contig 1). Twelve BAC-end sequences were isolated from seven clones and used to construct a restriction map. Two of them, 1118F and 1182R, matched different regions of the same gene, an *A. thaliana* kinesin-like protein (*A. thaliana* At3g20150.1). Comparison between the two *G. max* contigs and one *M. truncatula* contig identified six sequences conserved across all three contigs and another three between just one of the two *G. max* contigs and *M. truncatula* (Fig. 1*b*). Between *G. max* contig 1 and *M. truncatula* contig 1, six sequences and their homologs were organized in the same order, but the other two probes (RFLP A685 and BAC-end 1182R) revealed a case of rearrangement. Seven sequences were conserved between *G. max* contig 2 and *M. truncatula* contig 1. The relative order was conserved for three of them (5475F, 1180R, and 5477F/1120F), with the order of the other four undetermined.

Microsyntenic relationships within the RFLP B039 contig group

RFLP B039 probably represents an expressed protein, showing 99% identity over 278 bp to *G. max* TC138970 and 75% identity over 44 amino acids to a putative rice senescence-associated protein (AAL79714). B039 identified three contigs: *G. max* contig 1 (five BACs, on linkage group I), *G. max* contig 2 (two BACs), and *M. truncatula* contig 1 (two BACs) (Table 1). Twelve unique *G. max* BAC-end probes were isolated from the seven *G. max* BACs and four from the two *M. truncatula* BACs. The 12 *G. max* BAC-end probes were used to rescreen the *G. max* and *M. truncatula* BAC libraries. Twenty-six new *G. max* BAC clones were obtained, with nine belonging to contig 1 and 11 to contig 2. For the remaining six BAC clones that were identified, only one fragment cross-hybridized with a mixture of the 12 *G. max* probes. These 12 *G. max* probes did not identify additional *M. truncatula* contigs with microsynteny to *G. max*, apart from *M. truncatula* contig 1, so the four *M. truncatula* BAC-end probes were used to rescreen the *M. truncatula* libraries. Thirteen new *M. truncatula* BACs were identified, with three of them forming a second *M. truncatula* contig, contig 2. However, no BAC end from *G. max* and only three BAC ends from *M. truncatula* contig 1 hybridized to *M. truncatula* contig 2, so no restriction map was constructed for it. Restriction maps were constructed for the other three BAC contigs, resulting in 16 mapped BAC ends on *G. max* contig 1, 10 on *G. max* contig 2, and 19 on *M. truncatula* contig 1 (Fig. 2). Two adjacent *M. truncatula* BAC ends 0197F and 1167R matched different regions of *M. truncatula* TC52339 and the sequence of BAC-end 1106F was tandemly duplicated.

All 27 low-copy *G. max* sequences (26 BAC-ends plus RFLP B039) were used in cross-hybridizations between the two *G. max* contigs. Sixteen out of 22 sequences were conserved between a 158-kb portion of contig 1 from 7367R to 7366R (not including the two terminal sequences) and contig 2 (Fig. 2*a*). Of the 11 nonconserved sequences, five (4215R to 7367R, 7366R) were localized to the ends of contig 1 and presumably beyond the region of overlap. The relative order could be determined for four pairs of paralogs and was the same between the two contigs. For the remaining 12 conserved probes, between two and six adjacent probes from one of the *G. max* contigs cross-hybridized to the same restriction fragment(s) in both contigs.

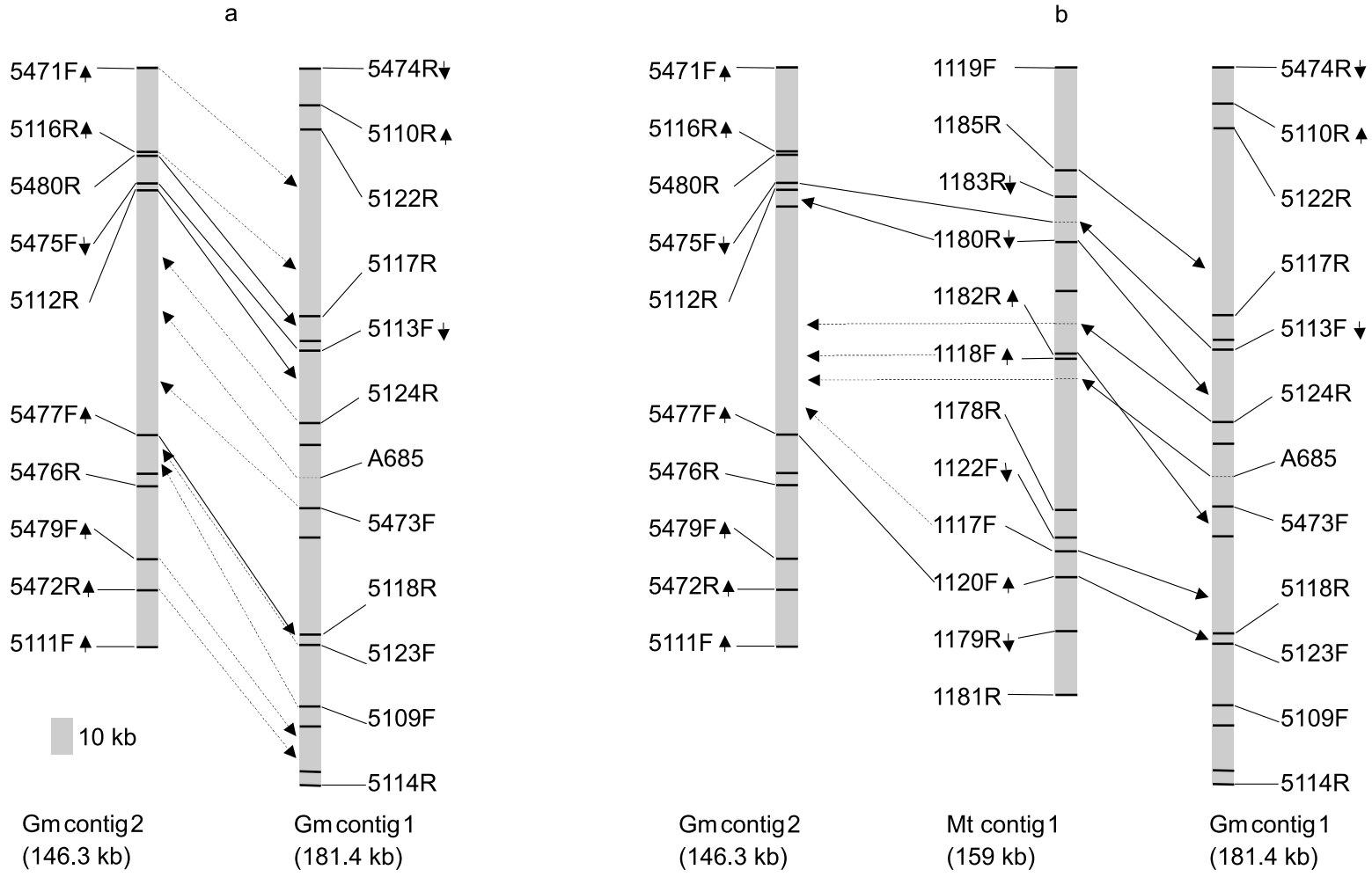
The two *G. max* contigs were compared with *M. truncatula* contig 1 by cross-hybridization and alignment of BAC-end sequences. A total of 17 different sequences were conserved between *G. max* and *M. truncatula* contig 1, while 11 were conserved across all three contigs (Fig. 2*b*). The other six were conserved between just one of the *G. max* contigs and *M. truncatula* contig 1. Of the 13 pairs of homologs identified in the 163.2-kb region of *G. max* contig 1 (7363R–7366R) and the 156.4-kb region of *M. truncatula* contig 1 (1172R–1170R), the relative order was identical for eight, while the order of the other five could not be determined. The transcription orientation was also determined for two pairs of homologs (1172R versus 7363R and 1164R versus 7333F) based on high sequence similarity, and in each case, both homologs had the same transcription orientation.

Microsyntenic relationships within the RFLP C063 contig group

Soybean RFLP C063 identified two contigs in *G. max*. One of them, designated contig 1, mapped to linkage group D1a+Q (Table 1). This contig contained 13 BAC clones, from which a 192.2-kb restriction map was constructed with 22 unique BAC ends plus RFLP C063 itself (Fig. 3*a*). The BAC-end probes were hybridized individually to digested *G. max* genomic DNA and all were classified as low copy. Ten of the 23 sequences were classified as coding, with two of them, 5238R and C063, matching different regions of the same gene (putative protein *A. thaliana* At5g13270.1). As contig 2 originally had only one BAC clone, all 22 BAC ends from contig 1 were used to rescreen the *G. max* BAC libraries. A total of 119 BAC clones were identified, with 49 out of 56 randomly chosen new clones confirmed by hybridization. Fingerprinting results assigned 27 BAC clones to contig 1 and the other 22 to contig 2. Nine BACs from contig 2 were sequenced at both ends, and 14 unique, low-copy BAC ends were isolated and physically mapped (Fig. 3*a*).

Between these two *G. max* contigs, there were four pairs of sequences that showed very high sequence similarity. Three pairs (7353R versus 5236R and 7325F versus 5238R/C063) were coding sequences with the same transcription orientation. The fourth pair included 7329F, a 746-bp sequence classified as coding with its 607- to 746-bp segment matching *G. max* TC133875, and 5247F, a 438-bp BAC end classified as noncoding. The remaining 18 probes from contig 1 and 10 from contig 2 were used as hybridization probes. In a 134.3-kb region of contig 1 (from 5236R to 5247F) and a 140.2-kb region of contig 2 (from 7353R to 7329F), all 19 probes, except 5231F and 7322F, cross-hybridized. Thus, 21 out of 23 different

Fig. 1. Microsyntenic relationship (*a*) between two *G. max* contigs anchored by *G. max* RFLP A685 and (*b*) between the two *G. max* contigs and *M. truncatula* contig 1. Restriction maps were built from fingerprinting and Southern blot hybridization results. Soybean BAC clones were digested with *Eco*RI and *M. truncatula* BAC clones with *Hin*dIII. Only those restriction sites that were used to define positions of BAC ends or regions showing cross-hybridization are indicated as horizontal lines across each gray bar. Although all *G. max* BACs were digested with *Eco*RI, *Hin*dIII restriction sites corresponding to the ends of BAC clones that were derived from *Hin*dIII-digested genomic DNA (Marek and Shoemaker 1997) are also indicated. If several BAC-end sequences from the same contig are identical or overlapping, only one is shown. Sequences connected by a solid line are homologs inferred from a high level of sequence identity; solid zlines with arrows indicate a probe cross-hybridizing to a syntenic contig, with the exact position of its homolog not determined; broken lines with arrows indicate adjacent probes cross-hybridizing to the same region in a syntenic contig, but with the exact positions and relative order of their homologs not determined; upward and downward arrows to the right of probe names indicate transcription orientations inferred from BLASTX results. The same designations also apply to Figs. 2 and 3. The exact position of probe A685 within the 5124R–5473F interval is arbitrary. (*a*) Homoeologous relationship between the two *G. max* contigs. Thirteen out of the 21 different sequences were conserved. The relative order was determined for four paralogs and was identical to that of the original BAC ends. (*b*) Syntenic relationship between the two *G. max* contigs and *M. truncatula* contig 1. Eight sequences were conserved between *G. max* contig 1 and *M. truncatula* contig 1, with six of them maintaining the same order. However, a small rearrangement involving sequences 1182R and A685 was observed. Between *G. max* contig 2 and *M. truncatula* contig 1, seven sequences were conserved.



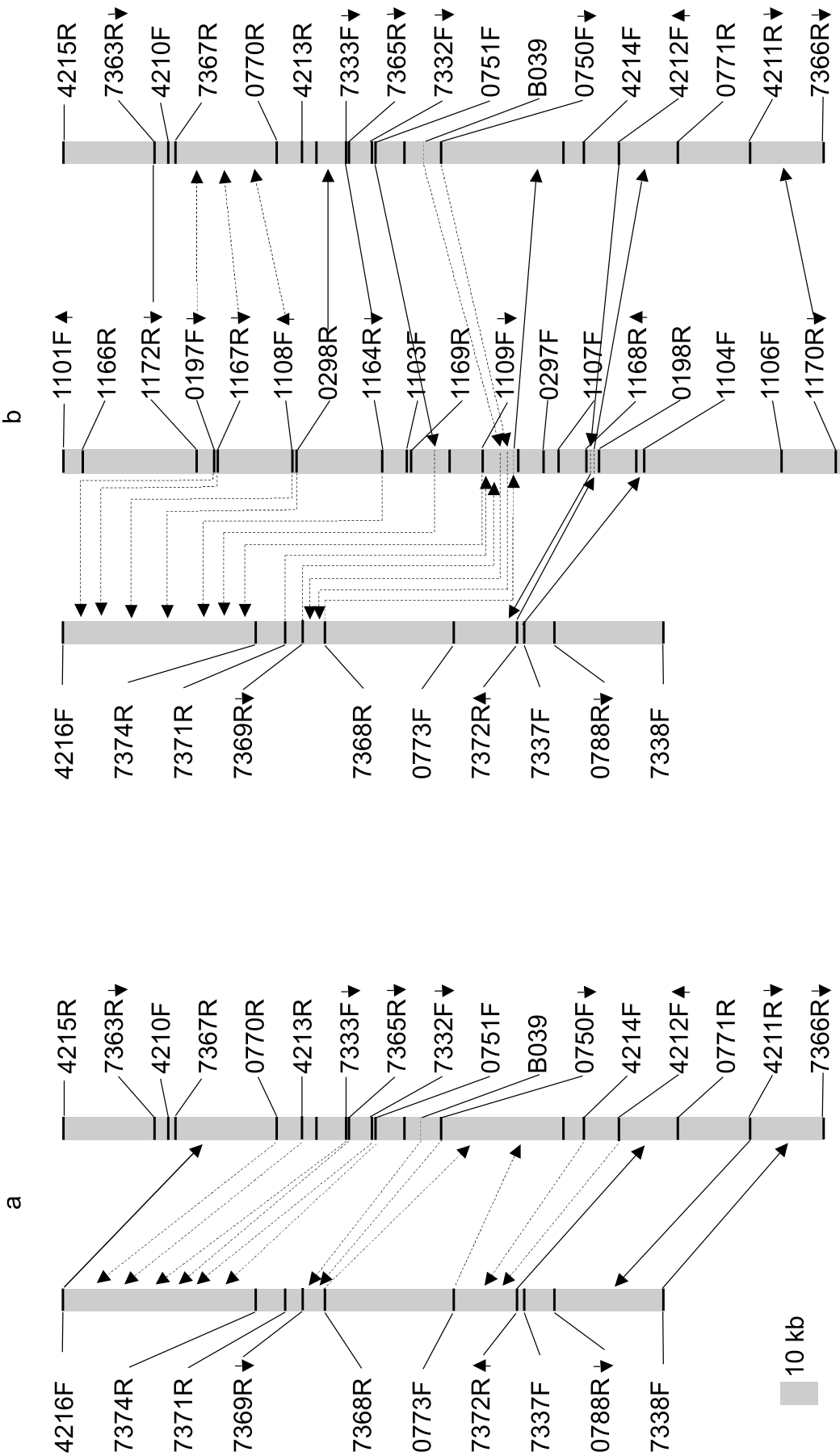


Fig. 2. Microsyntenic relationship (*a*) between two *G. max* contigs anchored by *G. max* RFLP B039 and (*b*) between the two *G. max* and one *M. truncatula* contigs. Anchor B039 sequenced was mapped to a 8.1-kb region between two restriction sites, as shown on contig 1, but its exact position and transcription orientation were not determined. (*a*) Homoeologous relationship between two *G. max* contigs. The lack of conservation for five sequences that mapped to either end of contig 1 is probably because contig 2 is 38.6 kb shorter than contig 1. (*b*) Syntenic relationships between the two *G. max* contigs and *M. truncatula* contig 1.

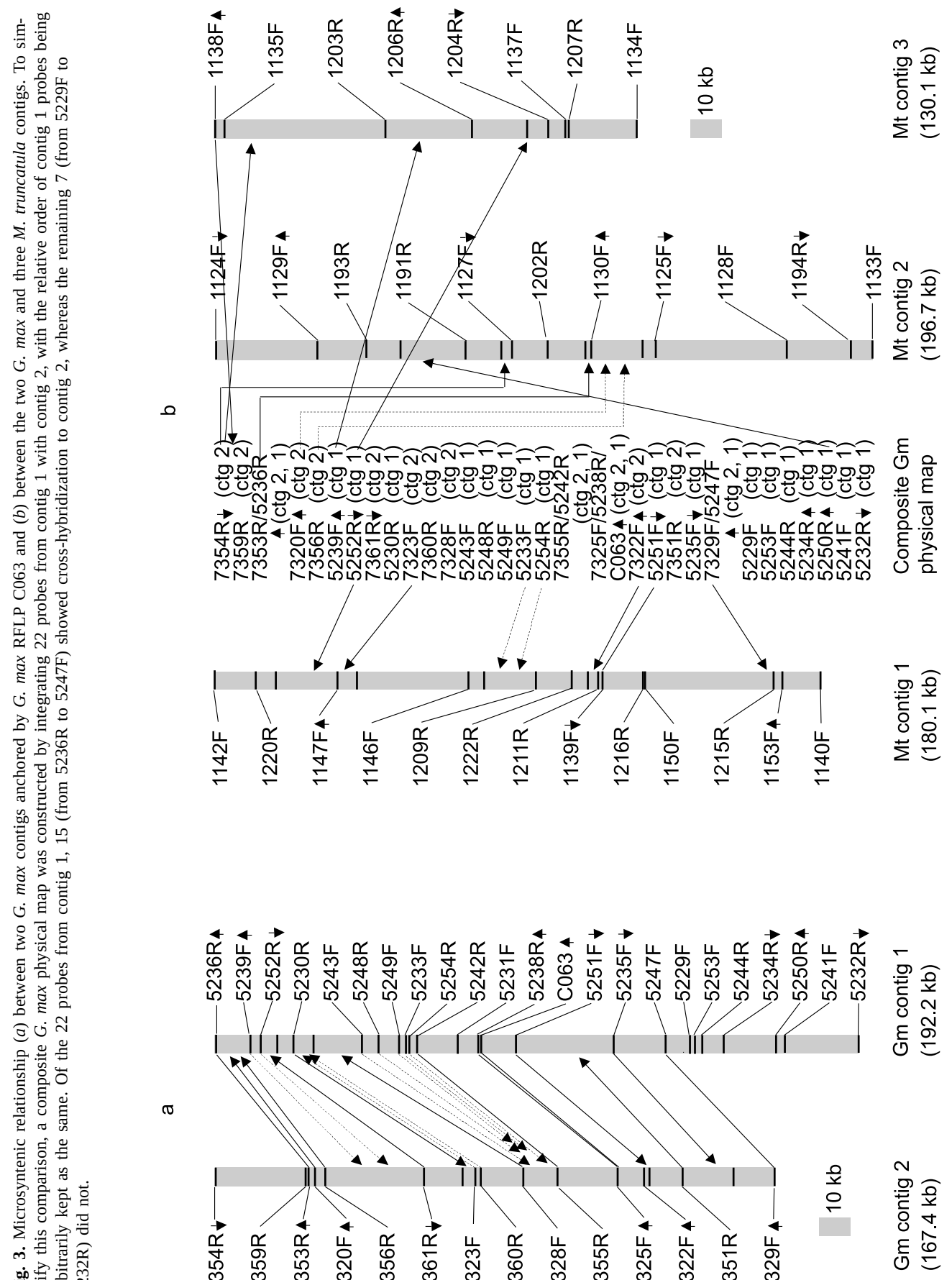
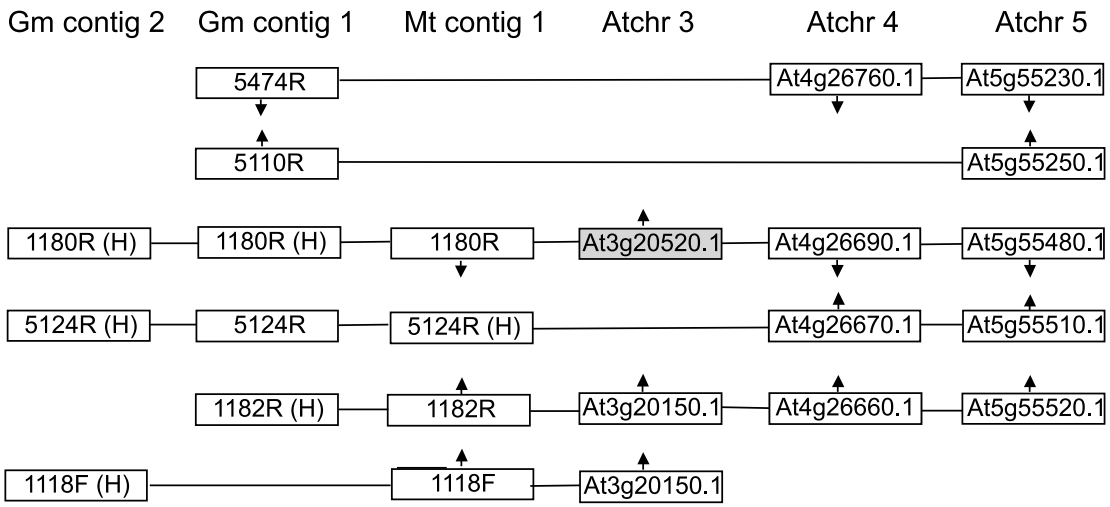


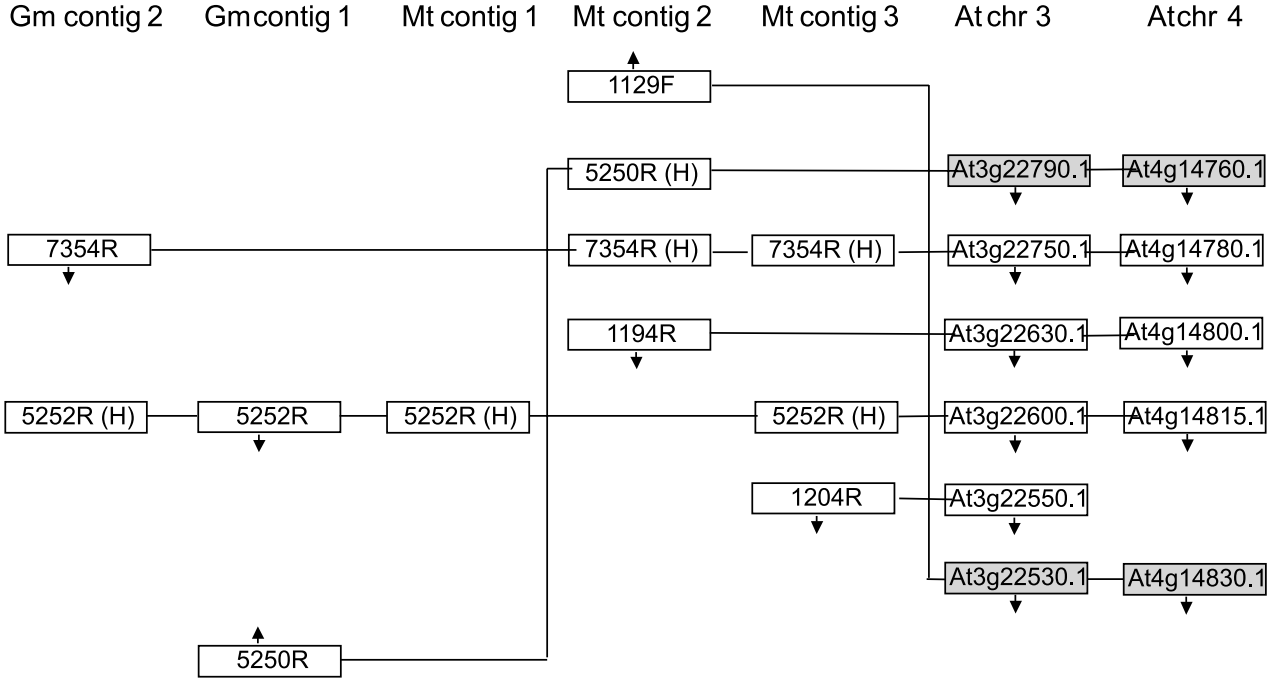
Fig. 3. Microsyntenic relationship (*a*) between two *G. max* contigs anchored by *G. max* RFLP C063 and (*b*) between the two *G. max* and three *M. truncatula* contigs. To simplify this comparison, a composite *G. max* physical map was constructed by integrating 22 probes from contig 1 with contig 2, with the relative order of contig 1 probes being arbitrarily kept as the same. Of the 22 probes from contig 1, 15 (from 5236R to 5247F) showed cross-hybridization to contig 2, whereas the remaining 7 (from 5229F to 5232R) did not.

Fig. 4. Graphic representation of the microsyntenic relationship among two groups of *G. max* and *M. truncatula* contigs and related regions in the *A. thaliana* genome. *Glycine max* and *M. truncatula* sequences were queried via tBLASTx against the *A. thaliana* genome sequence, and cases of microsynteny were inferred when homologs in *A. thaliana* were separated by a distance of less than 200 kb, each with an expected value of 1×10^{-8}. Arrow, transcription orientation; shaded square, *A. thaliana* sequence with inverted orientation relative to that of *G. max* and *M. truncatula*. "H" in parentheses indicates sequence homolog inferred from cross-hybridization or sequence alignment for which the transcription orientation was not determined. The number in the square indicates the gene model number (http://www.arabidopsis.org/). (*a*) Microsynteny between three A685-anchored *G. max – M. truncatula* contigs and three regions on *A. thaliana* chromosomes 3–5. Sequences 1118F and 1182R showed homology to a different region of the same gene, *A. thaliana* kinesin At3g20150.1. Six *G. max – M. truncatula* sequences, including five coding sequences, were conserved. (*b*) Microsynteny between six sequences from five C063-related *G. max – M. truncatula* contigs in two *A. thaliana* regions on chromosomes 3 and 4. Sequence 1204R, for which the best match in *A. thaliana* was the one on chromosome 3, was also included even though the expected value of 1×10^{-6} was slightly above the cutoff.

a



b



BAC-end sequences (91.3%) were conserved between these contigs (Fig. 3*a*). The other nine probes that did not show cross-hybridization were localized at either the upper end of contig 2 or the lower end of contig 1 and presumably lie outside the regions of overlap.

The relative order of paralogs could be determined for 12 of the 21 conserved sequences and was found to be identical between the contigs. For the remaining nine probes, between two and five adjacent probes cross-hybridized to corresponding regions on both contigs (Fig. 3*a*).

The 22 BAC ends from *G. max* contig 1 used to identify BAC clones in *G. max* contig 2 were also used to identify BACs in *M. truncatula*. A total of 73 *M. truncatula* BACs were identified and all were subjected to Southern hybridization. Of the 64 BACs confirmed as positives, 46 hybridized to two or more *G. max* BAC-end or RFLP probes. For the other 18 *M. truncatula* BACs, just one fragment hybridized to the mixture of *G. max* probes, so these 18 BACs were not analyzed further. Fingerprinting assigned the 46 BACs to three *M. truncatula* contigs (contigs 1, 2, and 3), with 21, 20, and five BAC clones, respectively. Ten BACs from *M. truncatula* contig 1, nine from contig 2, and four from contig 3 were sequenced at both ends, resulting in 13, 11, and 8 unique BAC-end sequences (Fig. 3*b*). Four *M. truncatula* BAC ends (1140F, 1202R, 1209R, and 1220R) each had a tandem duplicate nearby.

To simplify comparisons between the *G. max* and *M. truncatula* contigs, a composite *G. max* physical map containing 31 distinct BAC ends was constructed by integrating 22 probes from *G. max* contig 1 (except 5231F) and contig 2 (Fig. 1*b*). Fifteen probes were from a region between 5236R and 5247F that showed cross-hybridization to contig 2, and the other seven were from a region between 5229F and 5232R that did not show cross-hybridization and were apparently beyond the region of overlap. The relative order of probes from contig 1 was maintained.

Seven *G. max* sequences were conserved in *M. truncatula* contig 1, including six between both *G. max* contigs and *M. truncatula* contig 1 and one (7322F) between *G. max* contig 2 and *M. truncatula* contig 1 only (Fig. 3*b*). Six *G. max* syntenic probes were identified by cross-hybridization. The seventh, *G. max* BAC-end 5251F, showed strong sequence similarity to *M. truncatula* BAC-end 1139F. Homologs of five *G. max* sequences were organized in the same order as in *M. truncatula* contig 1, while the relative order of *M. truncatula* homologs for two adjacent *G. max* probes, 5233F and 5254R (only 0.9 kb apart on *G. max* contig 1), could not be determined. In total, this syntenic relationship involved a 121.2-kb region on *G. max* contig 1 (from 5252R to 5247F), a 134-kb region on *G. max* contig 2 (from 7356R to 7329F), and a 153.3-kb region on *M. truncatula* contig 1 (from 1220R to 1215R). Furthermore, when two BAC clones (5231F–5232R and 5235F–5236R) covering *G. max* contig 1 were hybridized to digested BACs from all three *M. truncatula* contigs, only BACs from *M. truncatula* contig 1 showed extensive cross-hybridization.

Five *G. max* BAC-end sequences were conserved in *M. truncatula* contig 2 (Fig. 3*b*). One of them, 7354R, had the same BLASTx hit as *M. truncatula* probe 1127F (*A. thaliana* expressed protein *A. thaliana* At3g22750.1). These four probes spanned 33.2 kb on *G. max* contig 2, with their homologs also residing in a region of less than 40-kb on *M. truncatula* contig 2. One instance of rearrangement was observed, involving *G. max* probe 5250R, which hybridized to a 3.9-kb fragment between *M. truncatula* BAC ends 1191R and 1193R.

Four sequences were conserved between *G. max* and *M. truncatula* contig 3, all located in a segment similar in position to the *G. max* region showing microsynteny to *M. truncatula* contig 2 (Fig. 3*b*). Homologs of three *G. max* probes (7354R, 5239F, and 5252R) maintained the same order in *M. truncatula* contig 3. *Medicago truncatula* BAC-end probe 1138F hybridized to a 5.1-kb fragment anchored by 7354R, representing one apparent case of rearrangement between these two regions.

Unlike the two homoeologous contigs in *G. max*, which were very highly conserved, only one or two sequences were conserved between any pair of *M. truncatula* contigs (not shown). When whole BAC clones representing each of the three *M. truncatula* contigs were used as probes, no cross-hybridization was observed.

Noncoding sequences are well conserved among *G. max* duplicates

We classified sequences as being coding or noncoding based on their similarity to protein or EST sequences in the databases and tested whether noncoding sequences are conserved between *G. max* duplicates. A total of 80 *G. max* probes were used in cross-hybridizations between homoeologous contigs from the three *G. max* contig groups examined. Seventeen of them mapped to contig ends for which no homoeologous segments were available for comparison. The remaining 63 sequences came from 27 coding and 36 noncoding sequences, with 75% (27/36) noncoding sequences conserved compared with 85% (23/27) coding sequences. Between *G. max* and *M. truncatula*, a total of 59 coding and 80 noncoding *G. max* and *M. truncatula* probes were used in cross-hybridization experiments. Twenty-seven coding (46%) and 13 noncoding (16%) probes were conserved, suggesting that coding sequences are about three times more likely to be conserved than noncoding sequences in comparisons between the genera. Notably, 13 of the 14 *G. max* and *M. truncatula* sequences later found to have microsynteny with *A. thaliana* (see below) were classified as coding sequences.

Multiple microsyntenic regions exist in the *A. thaliana* genome

A total of 139 low-copy BAC-end and RFLP sequences, including 32 from A685-related contigs, 44 from B039-related contigs, and 63 from C063-related contigs were searched against the *A. thaliana* genome sequence using TBLASTX. The criteria for inferring microsynteny were an expected value of $\leq 1 \times 10^{-8}$ for each hit, physical separation between homologs less than 200 kb in *A. thaliana*, and no more than five significant hits ($\leq 1 \times 10^{-8}$) for the sequence query. A total of 88 sequences (63.3%) did not have any matches and 18 (13.0%) only had matches less significant than 1×10^{-8}. Of the remaining 33 sequences with significant hits, six had more than five hits and were excluded from further analysis.

Ten *G. max* – *M. truncatula* sequences from the three A685-related contigs had five or fewer significant hits in *A. thaliana*. Five of these homologs were found in three homoeologous regions on *A. thaliana* chromosome 3, 4, or 5

Fig. 5. A model to describe the origin of the two *G. max*, three *M. truncatula*, and two *A. thaliana* genomic regions associated with RFLP C063. This model assumes that they are all descended from the same genomic segment in a hypothetical *G. max* – *M. truncatula* – *A. thaliana* ancestor. For simplification, only those probes that were conserved between at least two of the three genomes were included. The figure displays a model in which there were four total rounds of duplications leading to the present-day network of microsynteny, one in *G. max*, one in *A. thaliana*, and two in *M. truncatula*. The duplication in the *G. max* lineage that gave rise to *G. max* contigs 1 and 2 maybe quite recent, since both contigs still share substantial similarity with each other. There are two rather ancient duplications in the *M. truncatula* lineage, with the first one leading to contig 1 and the second one leading to contigs 2 and 3.

(Fig. 4*a*). *Glycine max* and *M. truncatula* contigs showed microsynteny to three *A. thaliana* regions, and all homologs were organized in the same order across the six regions. Among them, a region from 5474R to 1182R homolog on *G. max* contig 1 (about 118 kb) shared five conserved sequences with an 87-kb region on *A. thaliana* chromosome 5, implying a stronger conservation between the two regions. Just one of the homologs, At3g20520.1 on chromosome 3, was present in an inverted orientation.

Six sequences from the three B039-related *G. max* and *M. truncatula* contigs had five or fewer significant hits in *A. thaliana*. None of the regions could clearly be described as microsyntenic, as the *A. thaliana* hits were largely dispersed throughout the genome (data not shown). However, 1164R (7333F) and 7372R had homologs localized to a relatively small region of *A. thaliana* chromosome 4 (3171 bp away from each other), but the *A. thaliana* hit for 7372R was actually fourth out of the top five hits, with an expect value far less significant than the top three.

Eleven sequences from the five C063-related contigs had five or fewer significant hits in *A. thaliana*. Six had homologs in one or both of two microsyntenic regions identified on *A. thaliana* chromosomes 3 and 4, and all the homologs represented the best or top two hits (Fig. 4*b*). Sequence 1204R was also included, although its best hit in the *A. thaliana* genome (At3g22600.1) was not significant (1×10^{-6}). The 77 872-bp region on *A. thaliana* chromosome 3 contained all six *G. max* and *M. truncatula* sequence homologs, and the 35 472-bp region on chromosome 4 had five of them.

Of the five C063-related contigs, *M. truncatula* contigs 2 and 3 showed higher levels of overall conservation to the syntenic regions in *A. thaliana*. By contrast, only one sequence from *M. truncatula* contig 1 had homologs in both *A. thaliana* regions, even though it did show extensive microsynteny to *G. max* contigs 1 and 2. Sequence order was generally conserved, but two rearrangements were detected: 5250R on *G. max* contig 1 and 1129F on *M. truncatula* contig 2. It seems likely, therefore, that BAC-end 5250R has been transposed from somewhere close to BAC-end 5236R (Fig. 3*b*) to its current position on *G. max* contig 1 after the speciation of *G. max* and *M. truncatula*. Inverted orientations were observed for four *A. thaliana* homologs (Fig. 4*b*), all of which are associated with the rearrangements.

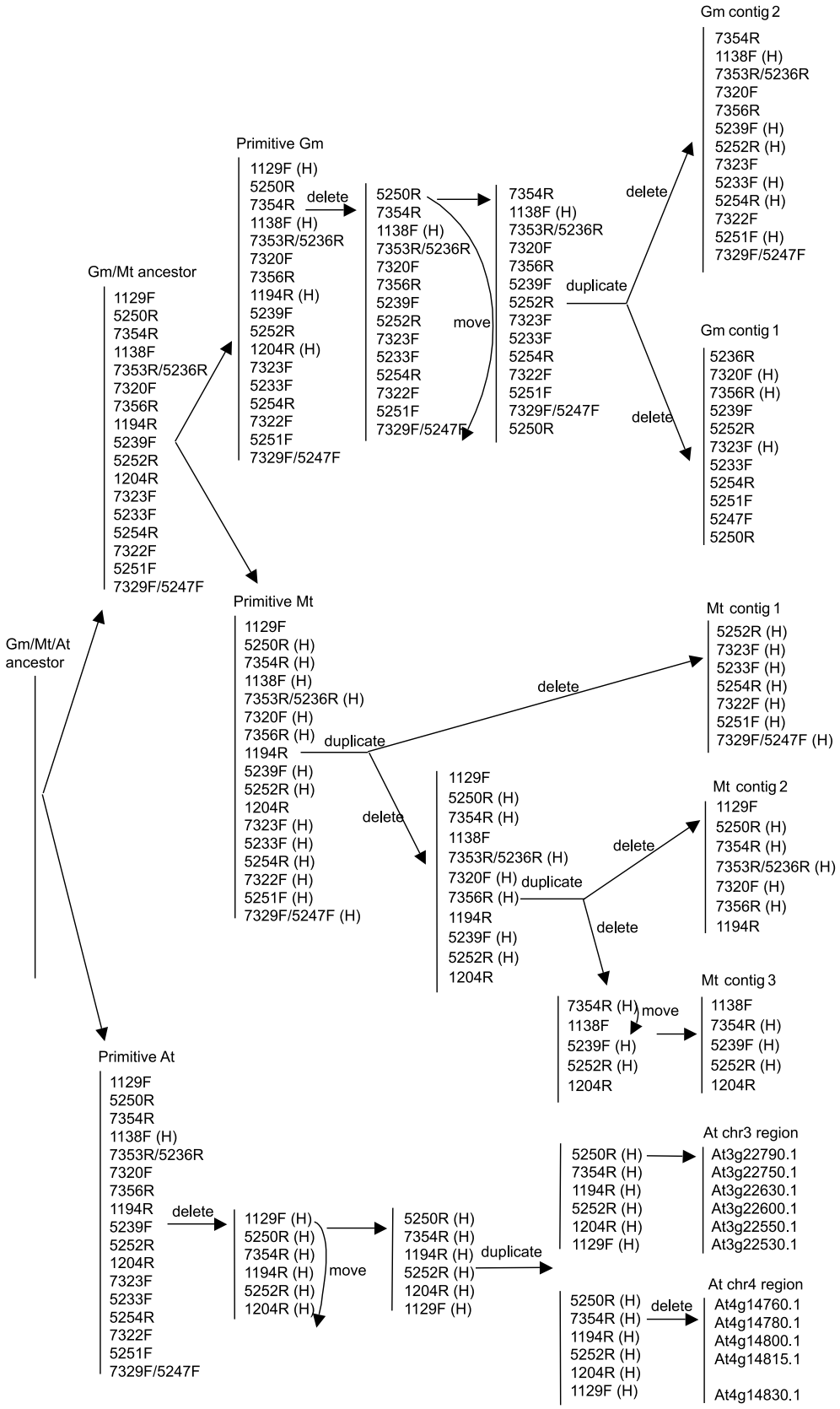
Discussion

Interrupted microsynteny among *G. max*, *M. truncatula*, and *A. thaliana*

Southern hybridization using BAC-end and RFLP probes revealed microsynteny between six of eight *G. max* contig groups and *M. truncatula* (Table 1). In particular, microsynteny between three *G. max* contig groups and *M. truncatula* appeared to be extensive, with 29% (range 22–39%) of all sequences tested being conserved at the level of cross-hybridization. Even between *G. max* – *M. truncatula* and *A. thaliana*, which diverged at least 90 million years ago (Lee et al. 2001), substantial microsynteny was still detectable for the A685 and C063 contig groups. It may be significant that these *G. max* and *M. truncatula* regions are associated with very low levels of repetitive DNA; only six out of 146 *G. max* and *M. truncatula* BAC-end probes were found by Southern hybridization to contain repetitive DNA or to show homology to retroelements. In fact, all of the *G. max* RFLP clones used to identify the contig groups in this study came from a *Pst*I genomic library (Keim et al. 1990) and were therefore more likely to be associated with hypomethylated and low-copy DNA (Burr et al. 1988). Available sequence data from grass species indicate that there are frequent gene-rich islands in plant genomes (Panstruga et al. 1998; Keller and Feuillet 2000), and a growing body of evidence suggests that this is also true for legumes (Kulikova et al. 2001; J. Mudge et al., in preparation). Therefore, these microsyntenic BAC contigs may come from gene-rich regions of the two genomes.

Of the three taxa examined, both *G. max* and *A. thaliana* genomes have apparently undergone multiple rounds of large-scale (even whole genome) duplications, and most of the resulting duplicates can still be detected by direct sequence comparison and (comparative) genetic mapping approaches (Shoemaker et al. 1996; Blanc et al. 2000; Simillion et al. 2002; Yan et al. 2003). We observed networks of microsynteny in all three cases, similar to what was described previously in tomato – *A. thaliana* (Ku et al. 2000) and *G. max* – *A. thaliana* sequence comparisons (Foster-Hartnett et al. 2002). A careful comparison among these contig groups in terms of number of conserved sequences, relative distance between adjacent sequences, sequence order, and transcription orientation enabled us to identify some contigs as being more highly conserved than others. For example, of the six syntenic regions associated with A685, five regions appear to have diverged relatively recently or diverged relatively slowly, with the exception of the region on *A. thaliana* chromosome 3 (Fig. 4*a*).

Despite the existence of microsynteny among *G. max*, *M. truncatula*, and *A. thaliana*, this microsynteny has always been interrupted in a variety of ways since speciation. Between *G. max* and *M. truncatula*, 99 of 139 sequences (71%) did not cross-hybridize between otherwise syntenic contigs, including 32 (out of 59) coding sequences. Homologs of these sequences may be present somewhere else in the genome, they may have been deleted in one of the two lineages, or they may have diverged to such an extent that homology is no longer detectable. A few instances of translocations were observed, although overall, 40 out of 139 sequences exhibited



unmistakable microsynteny. Such interrupted microsynteny has been reported between tomato and *A. thaliana* (Ku et al. 2000), between *G. max* and *A. thaliana* (Foster-Hartnett et al. 2002), and among several grass species (Bennetzen 2000; Devos and Gale 2000). As expected, the relationship between *G. max* – *M. truncatula* and *A. thaliana* is more distant.

Only 33 of the 139 *G. max* and *M. truncatula* sequences (24%) had significant matches in *A. thaliana* and only 13 of these 33 sequences showed microsynteny.

Evidence for a recent duplication in the *G. max* genome

Lee et al. (2001) compared sequences sampled from one pair of *G. max* duplicates with the *A. thaliana* genome sequence and found evidence indicating that the *G. max* genome may have undergone an ancient duplication. In previous work, we found that only two contigs from any *G. max* contig group tended to show extensive conservation, with a third (or any higher order) homoeologous contig showing much more limited conservation (Yan et al. 2003). This suggested the occurrence of one or more ancient duplications in *G. max* plus another much more recent one.

The present study extends these earlier results and indicates that contig-sized regions maintained very high levels of conservation in the three *G. max* contig groups that we examined in detail (Figs. 1*a*, 2*a*, and 3*a*), but much less so in the other five contig groups analyzed. In the three contigs with extensive microsynteny, 85% of coding and 75% of noncoding sequences showed conservation (although the number of noncoding sequences may have been overestimated, as a homology search alone would not be sufficient to identify all possible coding sequences). All sequence paralogs whose relative order could be determined were arrayed in the same order between the two *G. max* homoeologous contigs. Our previous investigation also indicated that at least 67% of the 37 *G. max* contig groups examined contain two contigs with extensive microsynteny (Yan et al. 2003). We therefore propose that the *G. max* genome has undergone a relatively recent large-scale (perhaps whole-genome) duplication, from which these pairs of homoeologous regions were derived.

Evidence for ancient duplications in *M. truncatula*

For the C063- and B039-anchored *G. max* contigs, two or three homologous contigs were identified in *M. truncatula*. In contrast with those in *G. max*, a much smaller proportion of probes showed cross-hybridization among these *M. truncatula* contigs. Southern hybridization using whole *M. truncatula* BAC clones was performed for 14 groups of *M. truncatula* contigs, and in all cases, no evidence for extensive cross-hybridization was observed (this study and unpublished results). Gualtieri and Bisseling (2002) compared a 70-kb *SYM2*-orthologous region on *M. truncatula* chromosome 5 and a 120-kb homoeolog located approximately 9 cM away and found that just four of 22 probes showed cross-hybridization. It seems likely, therefore, that these *M. truncatula* segments are the result of an ancient duplication event.

We identified seven C063-related syntenic regions from the three genomes: two in *G. max*, three in *M. truncatula*, and two in *A. thaliana*. The three *M. truncatula* contigs are highly diverged based on cross-hybridization results. *Medicago truncatula* contig 1 retained a higher level of homology to the two *G. max* contigs but seemed to lack microsynteny with the two *A. thaliana* regions (Fig. 4*b*). Instead, *M. truncatula* contigs 2 and 3 showed higher levels of microsynteny with the two *A. thaliana* regions. Therefore, it seems likely that *M. truncatula* contig 1 and *M. truncatula* contigs 2 and 3 resulted from one duplication event and *M. truncatula* contigs 2 and 3 resulted from the second duplication; both occurred early in the *M. truncatula* lineage. Based on these data, we proposed a model to explain the possible origin of these segments involving a total of four rounds of duplications (Fig. 5). The comparison among two *G. max* and two *M. truncatula* contigs related to RFLP probe B039 also supports this model. Soybean contigs 1 and 2 and *M. truncatula* contig 1 shared higher level of microsynteny, with 11 out of the 27 *G. max* sequences and eight out of the 19 *M. truncatula* sequences showing conservation (Fig. 2*b*). By contrast, no *G. max* sequence and only three *M. truncatula* sequences from contig 1 were conserved in *M. truncatula* contig 2. A plausible explanation is that there was an ancient duplication in *M. truncatula*. It is also possible that some ancient duplicates even predate the speciation of *G. max* and *M. truncatula*.

In summary, cross-hybridization and restriction mapping data both indicate very strong conservation between *G. max* homoeologous contigs but a much higher level of divergence between *M. truncatula* homoeologous contigs. The extent of microsynteny between *G. max* and *M. truncatula* varied considerably among the BAC contig groups and was accompanied by obvious sequence losses and rearrangements. These data further suggest a recent duplication in the *G. max* genome as well as ancient duplications in *M. truncatula*. These *G. max* and *M. truncatula* contigs are now being fully sequenced at the University of Oklahoma Genome Center (B. Roe, personal communication). Comparative sequence analysis of these regions will help us to estimate the time when duplication events may have occurred in both genomes. This may also help to reveal which genes have been selectively conserved or lost and what mechanism(s) are primarily involved in the maintenance of microsynteny that was observed.

Acknowledgments

We thank D. Larsen, R. Denny, S. Cannon, D. Danesh, and L.F. Marek for helpful discussions and for providing valuable materials. We also thank R. Staggs and T. Schmidt for assistance with some computational analysis. This research was supported by National Science Foundation grants DBI-0196179, DBI-0110206, and DBI 98-72565. This paper is published as part of a series from the Minnesota Agricultural Experiment Station.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. **215**: 403–410.

Bennetzen, J.L. 2000. Comparative sequence analysis of plant nuclear genomes: microcollinearity and its many exceptions. Plant Cell, **12**: 1021–1029.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell, **12**: 1093–1101.

Burr, B., Burr, R., Thompson, K., Albertson, M., and Struber, C. 1988. Gene mapping with recombinant inbreds in maize. Genetics, **118**: 519–526.

Cook, D.R. 1999. *Medicago truncatula* — a model in the making. Curr. Opin. Plant Biol. **2**: 301–304.

Danesh, D., Penuela, S., Mudge, J., Denny, R.L., Nordstrom, H., Martinez, J.P., and Young, N.D. 1998. A bacterial artificial chromosome library for *G. max* and identification of clones near

a major cyst nematode resistance gene. Theor. Appl. Genet. **96**: 196–202.

Devos, K.M., and Gale, M.D. 2000. Genome relationships: the grass model in current research. Plant Cell, **12**: 637–646.

Endre, G., Kereszt, A., Kevei, Z., Mihacea, S., Kaló, P., and Kiss, G.B. 2002. A receptor kinase gene regulating symbiotic nodule development. Nature (Lond.), **417**: 962–966.

Foster-Hartnett, D., Mudge, J., Larsen, D., Danesh, D., Yan, H.H., Denny, R., Penuela, S., and Young, N.D. 2002. Comparative genomic analysis of sequences sampled from a small region on soybean (*Glycine max*) molecular linkage group G. Genome, **45**: 634–645.

Gale, M.D., and Devos, K.M. 1998. Plant comparative genetics after 10 years. Science (Wash., D.C.), **282**: 656–659.

Gaut, B.S. 2002. Evolutionary dynamics of grass genomes. New Phytol. **154**: 15–28.

Grant, D., Cregan, P., and Shoemaker, R.C. 2000. Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between *G. max* and *Arabidopsis*. Proc. Natl. Acad. Sci. U.S.A. **97**: 4168–4173.

Gualtieri, G., and Bisseling, T. 2002. Microsynteny between the *Medicago truncatula SYM2*-orthologous genomic region and another region located on the same chromosome arm. Theor. Appl. Genet. **105**: 771–779.

Gualtieri, G., Kulikova, O., Limpens, E., Kim, D.J., Cook, D.R., Bisseling, T., and Geurts, R. 2002. Microsynteny between pea and *Medicago truncatula* in the SYM2 region. Plant Mol. Biol. **50**: 225–235.

Hall, A.E., Fiebig, A., and Preuss, D. 2002. Beyond the *Arabidopsis* genome: opportunities for comparative genomics. Plant Physiol. **129**: 1439–1447.

Keim, P., Diers, B., Olson, T., and Shoemaker, R.C. 1990. RFLP mapping in *G. max*: association between marker loci and variation in quantitative traits. Genetics, **126**: 735–742.

Keller, B., and Feuillet, C. 2000. Collinearity and gene density in grass genomes. Trends Plant Sci. **5**: 246–251.

Ku, H.-M., Vision, T., Liu, J., and Tanksley, S.D. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc. Natl. Acad. Sci. U.S.A. **97**: 9121–9126.

Kulikova, O., Gualtieri, G., Geurts, R., Kim, D.J., Cook, D.R., Huguet, T., de Jong, J.H., Fransz, P.F., and Bisseling, T. 2001. Integration of the FISH-pachytene and genetic maps of *Medicago truncatula*. Plant J. **27**: 49–58.

Lee, J.M., Grant, D., Vallejos, C.E., and Shoemaker, R.C. 2001. Genome organization in dicots. II. *Arabidopsis* as a 'bridging species' to resolve genome evolution events among legumes. Theor. Appl. Genet. **103**: 765–773.

Marek, L.F., and Shoemaker, R.C. 1997. BAC contig development by fingerprint analysis in *G. max*. Genome, **40**: 420–427.

Marek, L.F., Mudge, J., Darnielle, L., Grant, D., Hanson, N., Paz, M., Yan, H., Denny, R., Larson, K., Foster-Hartnett, D., Cooper, A., Danesh, D., Larsen, D., Schmidt, T., Staggs, R., Crow, J.A., Retzel, E., Young, N.D., and Shoemaker, R.C. 2001. Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. Genome, **44**: 572–581.

Marra, M.A., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D., and Waterston, R.H. 1997. High throughput fingerprint analysis of large-insert clones. Genome Res. **7**: 1072–1084.

Mayer, K., Murphy, G., Tarchini, R., Wambutt, R., Volckaert, G., Pohl, T., Dusterhoft, A., Stiekema, W., Entian, K.D., Terryn, N., Lemcke, K., Haase, D., Hall, C.R., van Dodeweerd, A.M., Tingey, S.V., Mewes, H.W., Bevan, M.W., and Bancroft, I. 2001. Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of *Arabidopsis thaliana*. Genome Res. **11**: 1167–1174.

Nam, Y.-W., Penmetsa, R.V., Endre, G., Uribe, P., Kim, D., and Cook, D.R. 1999. Construction of a bacterial artificial chromosome library of *Medicago truncatula* and identification of clones containing ethylene-response genes. Theor. Appl. Genet. **98**: 638–646.

O'Neill, C.M., and Bancroft, I. 2000. Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. Plant J. **23**: 233–243.

Panstruga, R., Buschges, R., Piffanelli, P., and Schulze-Lefert, P. 1998. A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. Nucleic Acids Res. **26**: 1056–1062.

Rossberg, M., Theres, K., Acarkan, A., Herrero, R., Schmitt, T., Schumacher, K., Schmitz, G., and Schmidt, R. 2001. Comparative sequence analysis reveals extensive microcollinearity in the lateral suppressor regions of the tomato, *Arabidopsis*, and *Capsella* genomes. Plant Cell, **13**: 979–988.

Salse, J., Piegu, B., Cooke, R., and Delseny, M. 2002. Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. Nucleic Acids Res. **30**: 2316–2328.

Shoemaker, R.C., Polzin, K., Labate, J., Specht, J., Brummer, E.C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J.P., Kochert, G., and Boerma, H.R. 1996. Genome duplication in soybean (*Glycine* subgenus *soja*). Genetics, **144**: 329–338.

Simillion, C., Vandepoele, K., Van Montagu, M.C.E., Zabeau, M., and Van de Peer, Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U.S.A. **99**: 13 627 – 13 632.

Tarchini, R., Biddle, P., Wineland, R., Tingey, S., and Rafalski, A. 2000. The complete sequence of 340 kb of DNA around the rice *adh1–adh2* region reveals interrupted collinearity with maize chromosome 4. Plant Cell, **12**: 381–391.

Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L., and Avramova, Z. 1999. Collinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc. Natl. Acad. Sci. U.S.A. **96**: 7409–7414.

Wong, G.K.S., Yu, J., Thayer, E.C., and Olson, M.V. 1997. Multiple-complete-digest restriction fragment mapping: generating sequence-ready maps for large-scale DNA sequencing. Proc. Natl. Acad. Sci.U.S.A. **94**: 5225–5230.

Yan, H.H., Mudge, J., Kim, D.-J., Larsen, D., Shoemaker, R.C., Cook, D.R., and Young, N.D. 2003. Estimates of conserved microsynteny among the genomes of *Glycine max*, *Medicago truncatula* and *Arabidopsis thaliana*. Theor. Appl. Genet. **106**: 1256–1265.